*Available online at* www.ijpab.com

DOI: http://dx.doi.org/10.18782/2320-7051.5046

**ISSN: 2320 – 7051**

*Int. J. Pure App. Biosci.* **6 (1):** 167-173 (2018)

*Review Article*

International Journal of Pure & Applied Bioscience

OPEN ACCESS

# RNA Sequencing (RNA-Seq): Method and Applications

**Bibha Rani***

Deptt. of Agril. Biotech. & Mol. Biology, Dr. Rajendra Prasad Central Agricultural University, Pusa, Bihar

*Corresponding Author E-mail: bibha9rani@gmail.com

Received: 24.06.2017 | Revised: 30.07.2017 | Accepted: 4.08.2017

**ABSTRACT**

*RNA sequencing or next generation sequencing (NGS) has emerged as a revolutionary tool in genetics, genomics, and epigenomics. Having high reproducibility and accuracy and cost effectiveness compared to other sequencing technologies, NSG has enabled genome wide investigations of various phenomena including allele specific expression, novel alternative splicing, single nucleotide polymorphism, copy number variants and novel transcripts. It also holds promise in discovering de novo transcription/splice junctions and small RNAs with high specificity. While RNA-Seq is a relatively new method, it has already provided unprecedented insights into the transcriptional complexities of a variety of organisms, including yeast, mice, Arabidopsis, and humans.*

***Key words:*** *RNA-seqencing, Transcriptome, cDNA libraries, Illumina technology.*

## INTRODUCTION

The genome is a store of biological information but on its own it is unable to release that information to the cell (http://www.nature.com/scitable/definition/genome-43). The initial product of genome expression is the transcriptome, the entire repertoire of transcripts in a species, represents a key link between DNA and phenotype whose biological information is required by the cell at a particular time (http://www.nature.com/scitable/definition/transcriptome-296).

RNA sequencing (RNA-Seq) is revolutionizing the study of the transcriptome. A highly sensitive and accurate tool for measuring expression across the transcriptome, it is providing visibility to previously undetected changes occurring in disease states, in response to therapeutics, under different environmental conditions and across a broad range of other study designs. RNA-Seq allows researchers to detect both known and novel features in a single assay, enabling the detection of transcript isoforms, gene fusions, single nucleotide variants, allele-specific gene expression and other features without the limitation of prior knowledge (www.illumina.com/techniques/sequencing/rna).

## MATERIALS AND METHODS

Library preparation is a key step of RNA-seq, because it determines how closely the cDNA sequence data reflect the original RNA population.

**Cite this article:** Rani, B., RNA Sequencing (RNA-Seq): Method and Applications, *Int. J. Pure App. Biosci.* **6(1):** 167-173 (2018). doi: http://dx.doi.org/10.18782/2320-7051.5046

The most straightforward approach is to simply synthesize double-stranded cDNA, to which the adapter can be ligated[1]. To prepare high quality cDNAs, it is important to start with a population of intact mRNAs. This is not always easy; mRNAs are very susceptible to cleavage by endogenous cellular ribonucleases, and some tissues or samples are very rich in these enzymes. Most eukaryotic mRNAs have several hundred bases of A at their 3' end. This poly A tail can be used to capture these mRNAs and remove contaminating rRNAs, tRNAs and other small cytoplasmic and nuclear RNAs. Unfortunately one also loses the fraction of mRNAs that lack a poly A tail. An oligo dT primer can be used with reverse transcriptase to make a DNA copy of the mRNA strand[2]. Alternatively, random primers can be used if one is searching for a particular mRNA or class of mRNAs. There are two general methods to convert RNA-DNA duplexes into cDNAs.

In first approach, the RNA strand is displaced or degraded, continue synthesis, after making a hairpin, until they have copied the entire DNA strand of the duplex. S1 nuclease can be used to cleave the hairpin and generate a cloneable end. Unfortunately, the S1 nuclease treatment can also destroy some of the ends of the cDNA. An alternative procedure is to use Rnase H to nick the RNA strands of the duplex. The resulting nicks can serve as primers for DNA polymerases like *Escherichia coli* DNA polymerase I. This eventually leads to a complete DNA copy except for a few nicks which can be sealed by DNA ligases.

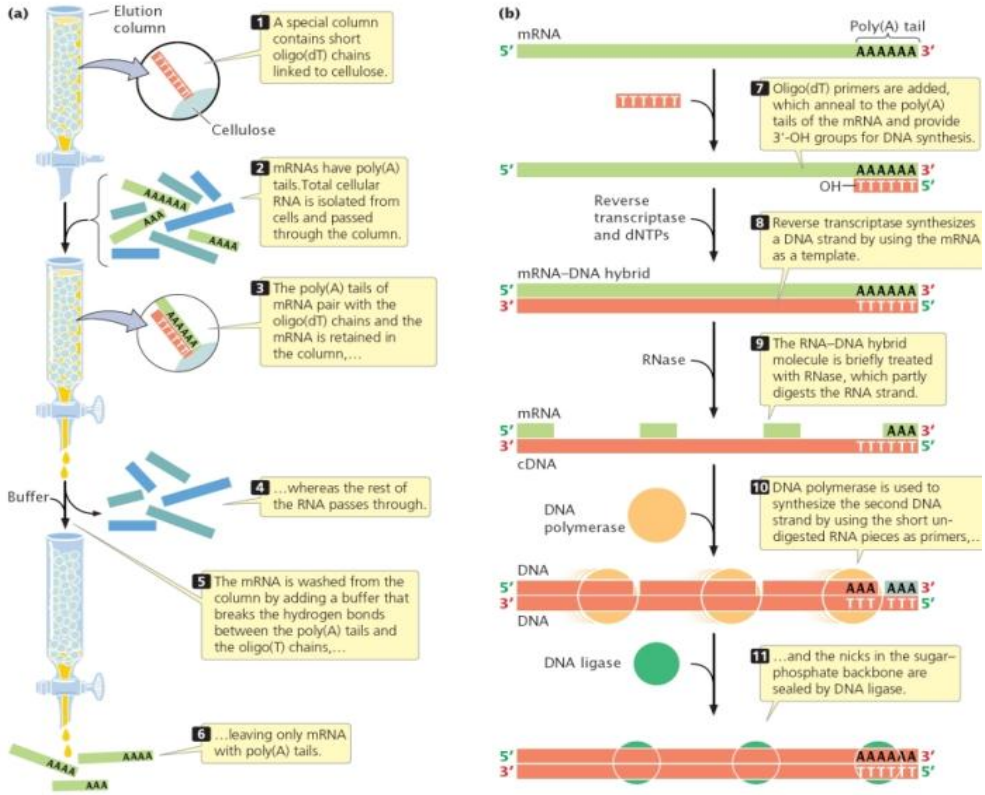


**Figure: Approaches to construction of cDNA libraries[3]**

Two experimental protocols for RNA-Seq are in common use: (a) single end and (b) paired end sequencing experiments. For single end experiments, one end (typically about 50 to 100 bp) of a long (typically 200 to 400 nucleotide) molecule is sequenced. For paired end experiments, typically 50 to 100 bp of both ends of a typically 200 to 400 nucleotide molecule are sequenced[4]. Using current Illumina technology, each time the sequencing machine is operated, eight samples (e.g., potentially eight different catalogues of gene

expression) can be interrogated (essentially) independently and tens of millions of reads are produced in each sample.

The advantages of RNA-Seq have enabled us to generate an unprecedented global view of the transcriptome and its organization for a number of species and cell types. Before the advent of RNA-Seq, it was known that a much greater than expected fraction of the yeast, *Drosophila melanogaster* and human genomes are transcribed and for yeast[5] and humans a number of distinct isoforms have been found for many genes[6]. However, the starts and ends of most transcripts and exons had not been precisely resolved and the extent of spliced heterogeneity remained poorly understood. RNASeq, with its high resolution and sensitivity has revealed many novel transcribed regions and splicing isoforms of known genes, and has mapped 5' and 3' boundaries for many genes.

**Advantage of RNA-Seq over other transcriptomics methods**

There are several advantages of RNA-Seq over other methods of transcriptome profiling. RNA-Seq can map transcribed regions and gene expression simultaneously while cDNA or EST sequencing is limited for gene expression. The dynamic range to quantify gene expression level is more than 8000-fold for RNA-Seq while just a few hundred compares to others. It has ability to distinguish different isoforms and allelic expression[7] (Table 1).

**Table 1: Comparison of RNA –Seq with other expression profiling methods**

| Transcriptomics methods | Principle | Resolution | Throughput | Reliance on genomic sequence |
|---|---|---|---|---|
| RNA-Seq | High-throughput sequencing | Single base | High | In some case |
| Microarray | Hybridization | From several to 100 bps | High | Yes |
| MPSS | Adaptor ligation and generation of 17 to 20 bps signature sequence | Single m-RNA transcript | High | Yes |
| SAGE | Generation of 14 bps tags | Single m-RNA transcript | High | No |

**RNA-Seq data analysis**

Once high-quality reads have been obtained, the first task of data analysis is to map the short reads from RNA-Seq to the reference genome, or to assemble them into contigs before aligning them to the genomic sequence to reveal transcription structure[8,9]. There are several programs for mapping reads to the genome, including ELAND, SOAP31, MAQ32, and RMAP (information about these can be found at the Illumina forum and at SEQanswers). However, short transcriptomic reads also contain reads that span exon junctions or that contain poly (A) ends, these cannot be analysed in the same way. For genomes in which splicing is rare (for example, *S. cerevisiae*) special attention only needs to be given to poly (A) tails and to a small number of exon–exon junctions. Poly (A) tails can be identified simply by the presence of multiple As or Ts at the end of some reads. Exon–exon junctions can be identified by the presence of a specific sequence context (the GT–AG dinucleotides that flank splice sites) and confirmed by the low expression of intronic sequences, which are removed during splicing. Transcriptome maps have been generated in this manner for

*S. cerevisiae*[9]. For complex transcriptomes, it is more difficult to map reads that span splice junctions, owing to the presence of extensive AS and *trans*-splicing. One partial solution is to compile a junction library that contains all the known and predicted junction sequences and map reads to this library. A challenge for the future is to develop computationally simple methods to identify novel splicing events that take place between two distant sequences or between exons from two different genes.

**Table 2: List of some open source solution for RNA-Seq data analysis**

| S/N | Software | Access address | Remarks |
|---|---|---|---|
| 1 | ArrayExpressHTS | http://bioconductor.org/packages/2.11/bioc/html/ArrayExpressHTS.html | allows preprocessing, quality assessment and estimation of expression of RNA-Seq datasets |
| 2 | easyRNASeq | http://www.bioconductor.org/packages/2.11/bioc/html/easyRNASeq.html | Calculates the coverage of high-throughput short-reads against a genome of reference and summarizes it per feature of interest (e.g. exon, gene, transcript) |
| 3 | Chipster | http://chipster.csc.fi/ | It contains over 300 analysis tools for next generation sequencing (NGS), microarray, proteomics and sequence data. Users can save and share automatic analysis workflows, and visualize data interactively using a built-in genome browser and many other visualizations. |
| 4 | ExpressionPlot | http://expressionplot.noip.me/cgi-bin/expressionplot/home.pl | Prepares raw sequencing or Affymetrix Array data and a web based front end which offers a biologically centered to browse, visualize and compare different data sets |
| 5 | GenePattern | http://www.broadinstitute.org/cancer/software/genepattern/modules/RNA-seq/ | GenePattern offers a set of tools to support a wide variety of RNA-seq analyses, including short-read mapping, identification of splice junctions, transcript and isoform detection, quantitation, differential expression, quality control metrics, visualization, and file utilities |
| 6 | NGS-Trex | http://www.ngs-trex.org | Allows the user to upload raw sequences and easily obtain an accurate characterization of the transcriptome profile after the setting of few parameters required to tune the analysis procedure. The system is also able to assess differential expression at both gene and transcript level |
| 7 | RobiNA | http://mapman.gabipd.org/web/guest/robin | Offers a variety of quality control methods that can be used to gain an overview of the experimental data technical quality and structure. Using the built-in graphical experiment designer, researchers can define which samples to compare. |

**Recent advances in RNA-seq methods**

Recently developed approaches allow more comprehensive understanding of transcription initiation sites, the cataloguing of sense and antisense transcripts, improved detection of AS events and the detection of gene fusion transcripts, which has become increasingly important in cancer research. It also allows the selection of specific RNA molecules before RNA-seq, allowing transcriptomics studies with more focused aims[10, 11].

**a. Mapping transcription start site**

The mapping of transcription start sites (TSSs) at nucleotide resolution is necessary to fully define RNA products and to identify adjacent promoter regions that regulate the expression of each transcript. This involved sequencing of cloned cDNA products derived from RNAs with intact 5' ends (for example, containing a 5' cap structure). Although useful, the technology required high quantities of input RNA and generated only short reads (~20 nucleotides) per TSS[12,13].

**b. Strand-specific RNA-seq**

Standard RNA-seq approaches generally require double-stranded cDNA synthesis, which erases RNA strand information. In addition, during first-strand cDNA synthesis, spurious second-strand cDNA artefacts can be introduced, owing to the DNA-dependent DNA polymerase (DDDP) activities of reverse transcriptases, and can confound sense versus antisense transcript determination. The strategies that have been developed to generate strand-specific information generally rely on one of three approaches. The first involves the ligation of adaptors in a predetermined orientation to the ends of RNAs or to first-strand cDNA molecules. The known orientations of these adaptors are used as reference points to obtain RNA strand information. A second approach is the direct sequencing of the first-strand cDNA products that are generated, either in solution or on surfaces. Last, a third approach is the selective chemical marking of the second-strand cDNA synthesis products or RNA. These strategies have already begun to contribute to our understanding of transcriptomes, including mapping of translation states of RNAs (for example, polysome profiling) and identification of novel promoter-associated RNAs[14, 15].

**c. Characterization of alternative splicing patterns**

Given the importance of AS patterns in development and the fact that 15 to 60% of known disease-causing mutations affect splicing, it will be crucial to catalogue the complete repertoire of splicing events and to understand how altered splicing patterns contribute to development, cell differentiation and human disease. Initial splice-site mapping studies using RNA sequencing-based approaches were limited by read length, which prevented the reliable alignment to the genome of the two independent exonic portions of each read, representing the exon splicing event. Improvements to current sequencing technologies now enable longer read lengths, allowing better mapping of the reads to the alternatively spliced exons. This improvement comes from being able to partition the reads into multiple pieces and to align each piece independently to the genomes[16, 17].

**d. Gene fusion detection**

RNA-seq combined with computational analyses, analogous to the ones described earlier for splice-site detection, can also be used to identify gene fusion events in disease tissues, which have particular importance for cancer research. DNA-based approaches cannot identify fusion events that are due to non-genomic factors, such as *trans*-splicing and read-through events between adjacent transcripts. Paired-end RNA-seq can be particularly advantageous for fusion identification because of the increased physical coverage it offers. This approach has led to important biological findings in oncology, offering potential targets for therapeutic modulation[18, 19].

## CONCLUSION

RNA-seq can provide complete transcriptional characterization of all the cells of an organism. This technique produces much significant information on how the

transcriptome deployed in different cell types and tissues, how gene expression changes across development states and how it varies within and between species. Sequencing transcripts (that is, expressed genes) is inherently cheaper than sequencing genomes, because it obviates the need to sequence the intronic and intergenic regions, which can be orders of magnitude larger. From this information one can generate new hypotheses about biology or test existing ones. With its unprecedented throughput, scalability, and speed, next-generation sequencing enables researchers to study biological systems at a level never before possible.

## REFERENCES


1. He, Y., Vogelstein, B., Velculescu, V.E., Papadopoulos, N., Kinzler, K.W. The antisense transcriptomes of human cells. *Science* **322:** 1855–1857 (2008).
2. Ingolia, N.T., Ghaemmaghami, S., Newman J.R.S., Weissman J.S. Genome-wide analysis *in vivo* of translation with nucleotide resolution using ribosome profiling. *Science* **324:** 218–223 (2009).
3. Benjamin, P. (2005) Genetics: A Conceptual Approach 2nd edition (Adapted from nature education).
4. Wang, Z., Gerstein, M., Snyder, M. RNA-Seq: a revolutionary tool for transcriptomics. *Nat Rev Genet*. **10(1):** 57–63 (2009).
5. Velculescu, V.E., Zhang, L., Zhou, W., Vogelstein, J., Bassett Jr., D.E., Basrai, M.A., Hieter, P., Vogelstein, B., Kinzler, K.W. Characterization of the Yeast Transcriptome. *Cell* **88:** 243-251 (1997).
6. Sorek, R. and Ast, G. Intronic sequences flanking alternatively spliced exons are conserved between human and mouse. *Genome Res.* **13:** 1631–1637 (2003).
7. Giannoukos, G., Ciulla, D.M., Huang, K., Haas, B.J., Izard, J., Livny, J., Earl, A.M., Gevers, D., Ward, D.V., Nusbaum, C., Bruce, W.B., Levin, J.Z. ,Gnirke, A. Efficient and robust RNA-seq process for cultured bacteria and complex community transcriptomes. *Genome Bio*.**13:** 1-13 (2012).
8. Jiang, H. and Wong, W.H. Statistical inferences for isoform expression in RNA-Seq. *Bioinfo*. **25(8):** 1026-1032 (2009).
9. Mortazavi, A., Williams, B.A., McCue, K., Schaeffer, L., Wold, B. Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat Methods* **5**:621–628 (2008).
10. Bryant, D.W. Jr, Priest, H.D., Mockler, T.C. Detection and quantification of alternative splicing variants using RNA-seq. *Methods Mol Biol*. **883:** 97-110 (2012).
11. Aschoff, M., Hotz-Wagenblatt, A., Glatting, K.H., Fischer, M., Eils, R., König, R. SplicingCompass: differential splicing detection using RNA-seq data. *Bioinfo*. **29(9):** 1141-1149 (2013).
12. Cortes, T., Schubert, O.T., Rose, G., Arnvig, K.B., Comas, I., Aebersold, R., Young, D.B. Genome-wide mapping of transcriptional start sites defines an extensive leaderless transcriptome in Mycobacterium tuberculosis. *Cell Rep*. **5(4):** 1121-1152 (2013).
13. Ozsolak, F., Milos, P.M. RNA sequencing: advances, challenges and opportunities. *Nat. Rev. Genet.* **12(2):** 87-98 (2011).
14. Levin, J.Z., Yassour, M., Adiconis, X., Nusbaum, C., Thompson, D.A., Friedman, N., Gnirke, A., Regev, A. Comprehensive comparative analysis of strand-specific RNA sequencing methods. *Nat Methods* **7(9):** 709-715 (2010).
15. Borodina, T., Adjaye, J., Sultan, M. A strand-specific library preparation protocol for RNA sequencing. *Methods Enzymol.* **500:** 79-98 (2011).
16. Florea, L., Song, L., Salzberg, S.L. Thousands of exon skipping events differentiate among splicing patterns in sixteen human tissues. *F1000Research*. **2:** 188 (2013).
17. Pedro, G. F., Rico P.J.D. Transcriptome characterization by RNA sequencing identifies a major molecular and clinical subdivision in chronic lymphocytic

leukemia. *Genome Res*. **24:** 212-226 (2013).

18. Trapnell, C., Williams, B.A., Pertea, G., Mortazavi, A., Kwan, G., Van Baren, M.J., Salzberg, S.L., Wold, B.J., Pachter, L. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat Biotechnol*. **28(5):** 511-515 (2010).
19. Treangen, T.J. and Salzberg, S.L. Repetitive DNA and next-generation sequencing: computational challenges and solutions. *Nat. Rev. Genet*. **13(1):** 36-46 (2012).